Instruction Steps of Bioinformatic Analysis

Bioinformatic analysis involve 3 major sections:

SECTION 1: SEQUENCING DATA ANALYSIS

The sequencing quality was evaluated by the Illumina Sequencing Analysis Viewer and FastQC software (Babraham Bioinformatics). Sequencing adapters and 3′ low quality bases were trimmed from raw sequencing reads using a custom algorithm and then aligned to the C➔T *in silico* converted hg19 reference genome, using Bismark (Bowtie2 as the default aligner behind Bismark). Aligned reads were then evaluated by Picard, for metrics that measured the performance of target capture based bisulfite sequencing assays (http://broadinstitute.github.io/picard).The biases of specific motifs or GC-enriched regions were excluded. After the preliminary analysis, we calculated the average coverage as well as the missing rate for each CpG site. The CpG sites with coverage less than 30x and/or with missing rate greater than 0.20 were filtered out. On-target rate, coverage (region of interest; on target), duplication rate, conversion rate and some other parameters were extracted as indicators of NGS QC (Quality Control).

SECTION 2: DIFFERENTIAL ANALYSIS

In order to determine the differences between groups (malignant vs benign, and these tests were done by each cancer type individually) in a statistical way, t-test (“t.test” function from programming language of R), and DSS test (Differential methylation signature analysis, identifying differentially methylated CpG sites by comparing samples from two groups using package DSS from programming language of R) were applied. Both their results were with false discovery rate (fdr) adjustment for the p-value (threshold used here was 0.05).

SECTION 3: MODEL CONSTRUCTION & ANALYSIS

Linear-based logistic regression model (“LR”, using “LogisticRegression” function in “sklearn” package from programming language of python) and tree-based random forest model (“RF”, using “RandomForestClassifier” function in “sklearn” package from programming language of python) was constructed using methylation markers (single site CpGs) as input features and pathology classes as label (digitalize the labels using 1 for malignant, 0 for benign/normal). The performance of the model was evaluated with Receiver Operator Characteristic (ROC) curve and Area Under Curve (AUC) values.